Age, Gender and Ethnicity Detection

K. Jayakar Sai Sushanth Reddy (18CS01004)
K. Kishorereddy (18CS01007)

Objectives of the Project:

Analyze the Dataset and build Machine Learning models to detect age, gender and ethnicity of a given facial image using Classification techniques. Evaluate and compare the accuracies of the models used and arrive at a conclusion.

Data Source:

The dataset is obtained from Kaggle at the following link https://www.kaggle.com/nipunarora8/age-gender-and-ethnicity-face-data-csv. It includes a CSV of facial images that are labelled on the basis of age, gender, and ethnicity. The dataset includes 27305 rows and 5 columns.

Attributes:

Attribute Name	Description	Attribute Type	Attribute Values
age	Age of the person in the image	Continuous, Integer	Range (1, 116)
gender	Ethnicity of the person	Categorical, Integer	{0, 1}
ethnicity	Gender of the person	Categorical, Integer	{0, 1, 2, 3, 4}
img_name	Name of the image	String	*.jpg
pixels	Array to String of the image pixels	String	String of 2304 space separated pixel values

Data Pre-Processing:

- The img_name attribute was dropped since it was simply the file names of the images used when the data set created and does not signify as a training attribute.
- For the target variables age, gender and ethnicity, the corresponding columns in the dataset were selected.
- For the input variables, the strings in 'pixels' column of the dataset were split into 2304 columns of pixel values, each row representing an image of 48*48 pixels.

Cleaning:

Chosen Dataset was already clean and no further operations were required.

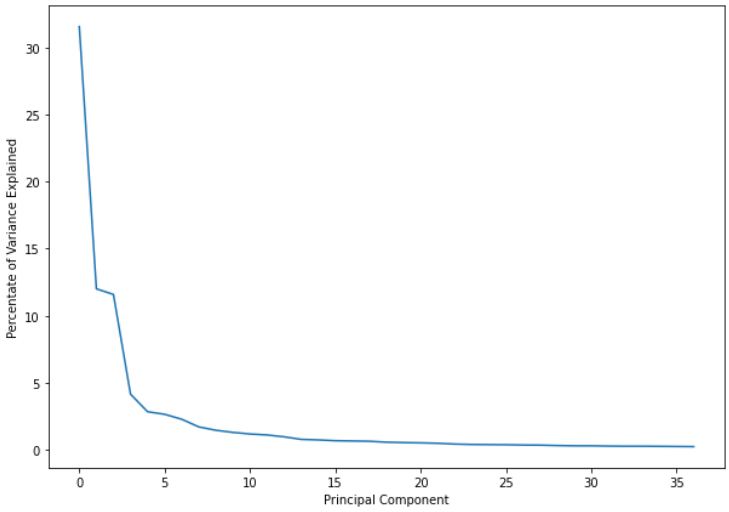
Data Transformation:

Z-Score Normalization was applied on the 2304 columns of pixel values.

For the CNN model [0-1] normalization was used as we knew the bounds of the pixel values.

Data Reduction:

Principal Component Analysis (PCA) was applied replacing 2304 input variables by 37 PCs which explained 85% of the variance.

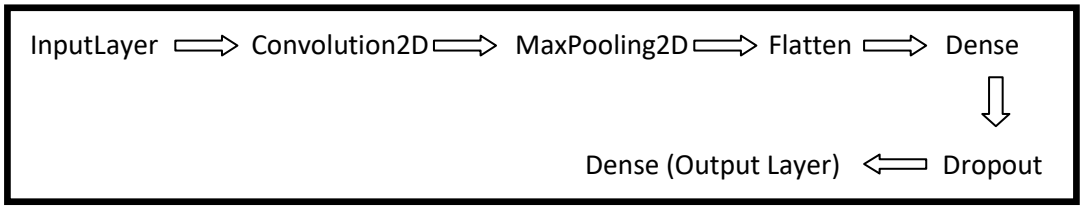


Scree Plot

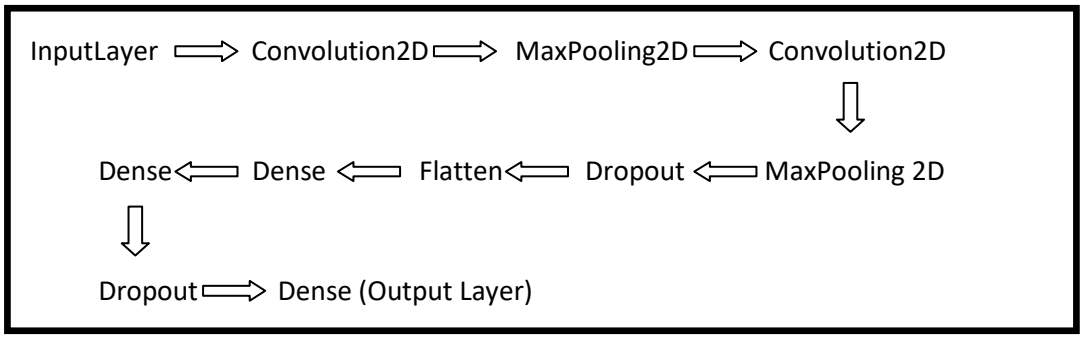
Methods:

Following methods were used in building models:

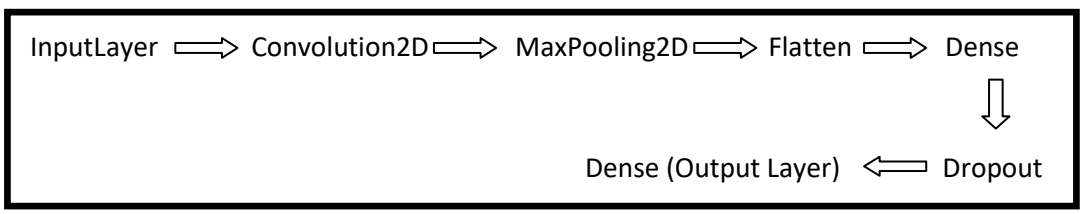
- Gaussian Naïve Bayes:
  - Since the standardized P.C's of pixels form a continuous variable, a basic method like Gaussian Naive bayes algorithm was used to train each of the Age, Ethnicity and Gender classification models.
- Logistic Regression:
  - Standard Logistic Regression is one of the basic models used in classification problems, its more advanced than the Gaussian NB method, hence as an experiment Logistic regression was used to train the Age, Ethnicity and Gender models and the performances were observed.
- Convolutional Neural Networks
  - Convolutional Neural Networks observes and recognizes patterns from the input data and uses the information in classifying inputs unlike the Gaussian NB and Logistic regression methods which rely completely on individual pixel values and don't leverage any patterns from the data.
  - Softmax Activation was used in case of Classification models (Gender and Ethnicity classification) with no. of neurons = no. of classes.
  - In regression case (Age Detection) the output layer consisted of only one neuron with relu activation.
  - Model Architecture for Age Detection:



  - Model Architecture for Gender Detection:



  - Model Architecture for Ethnicity Detection:



Results and Analysis:

Model	Age	Ethnicity	Gender
Gaussian NB	0.087 **(P.C.A – 0.9)**	0.544 **(P.C.A – 0.95)**	0.770 **(P.C.A – 0.9)**
Logistic Regression	0.131 **(P.C.A – 0.9)**	0.708 **(P.C.A – 0.95)**	0.827 **(P.C.A – 0.9)**
CNN	R2 Score: 0.630 MSE: 145.177 MAE: 8.955	Accuracy: 0.765	Accuracy: 0.880

Confusion Matrices for CNN models:

Gender:

Predicted Actual	0	1
0	3620	483
1	456	3264

Ethnicity:

Predicted Actual	0	1	2	3	4
0	3035	81	101	145	13
1	208	1160	30	64	12
2	188	29	871	25	9
3	285	73	28	887	23
4	350	47	28	96	35

- Gaussian NB and Logistic regression can be trained faster than CNN models.
- In case of Gender Classification problem the results from these models are satisfactory given the computation power required for training, while for ethnicity and age detection problems since there are many classes the performance decreases.
- We have observed that the Convolutional Neural Network models outperform the Gaussian NB and the logistic regression classification model in all the problems.

Appendix:

The code has been uploaded at Click here